**Great Learning**

**Capstone Project – Final Report (Milestone 2)**

**Pneumonia Detection Challenge**

**Group:** **CV Group 4**

**Authors:**

1. **Ajay Devnani**
2. **Manojkumar Sivaraman**
3. **Naveen Kumar Kaginelli**
4. **Sumeet Kumar**
5. **Umesh Singh Kushwaha**

**Abstract**

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it and the infection causes inflammation in the air sacs in your lungs, which are called alveoli. The alveoli fill with fluid or pus, making it difficult to breathe. Typically, X-ray helps doctors to look for signs of inflammation or opacities in chest which when present can indicate the Pneumonia infections. Since Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally, it is crucial to identify and react swiftly if there are any infections identified.

**Business Perspective**: As the detection and reaction time is vital and the infection is detected using the X-ray Images, Image processing techniques can be leveraged from the emerging AI technology on these images to predict the presence of opacities. Powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making. This will also assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (e.g., radiology).

To achieve the goal, techniques such as instance segmentation (Mask RCNN) and semantic segmentation are used to create the Pneumonia prediction model which can predict Pneumonia and also the position of lung inflammation (lung opacities).

**Problem Statement**

The problem is about detecting lung inflammations (opacities) corresponding diagnosis of Pneumonia on chest radiographs (images). Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image. In the data, some of these such area labeled as “Not Normal No Lung Opacity”.

All lung opacities may not attribute to Pneumonia as the Pneumonia is one of the several diseases that can occur on a chest (lungs) radiograph. The “Not Normal No Lung Opacity” class indicates that, while pneumonia determined not to be present, there could be nonetheless some type of abnormality on the image. And oftentimes this finding may mimic the appearance of true pneumonia. A radiograph may contain one or more than one locations for any possible Pneumonia case.

**Summary of data and findings:**

The data is spread across different files and folders. The details are as given below,

* **stage\_2\_train\_images:** It contains a set of raw medical images (DICOM files) for training models. The DICOM files contain a combination of header metadata as well as underlying raw image arrays for pixel data.
* **stage\_2\_test\_images:** It contains a set of raw medical images (DICOM files) for testing the model. The file contains a combination of header metadata as well as underlying raw image arrays for pixel data**.**
* **stage\_2\_train\_labels.csv:** This CSV file contains detailed information about the labels (Patient ID, bounding boxes for lung opacity and target 1 or 0 indicate the presence of abnormality i.e. Pneumonia)
* **stage\_2\_detailed\_class\_info.csv:** This CSV file contains information regarding three possible classes in the data - namely normal, lung opacity and no lung opacity (not normal).
* **DICOM files:** The original medical images are stored in a special format called DICOM files (\*.dcm). It contains a combination of header metadata as well as underlying raw image arrays for pixel data.

**Findings:**

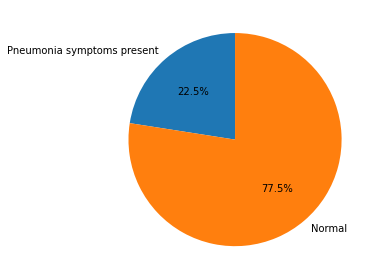
* The most important thing is that all lung opacities may not attribute to pneumonia, as the pneumonia is one of the several diseases that can occur on a chest radiograph.
* A radiograph may contain one or more than one bounding boxes for any possible pneumonia case.

**EDA and Pre-processing**

**Exploratory data analysis (EDA):**

1. **Distribution of Pneumonia Vs Non-Pneumonia:**

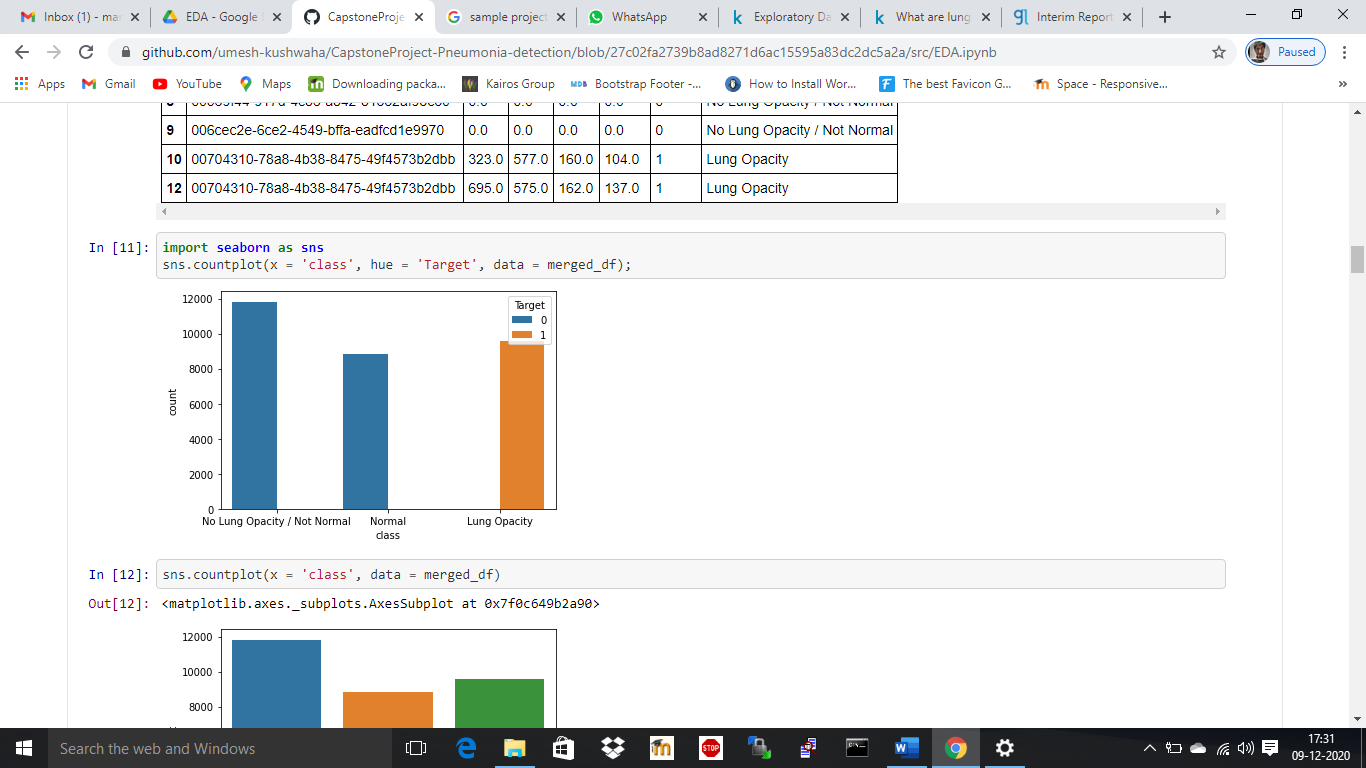
* 33.2% (8,851) patients are normal, do not have any lung related abnormalities
* 22.5% (6012) patients have *lung opacities* which attributes to pneumonia.
* 44.3% (11,821) patients do not have pneumonia but are not normal possibly due to other lung ailments.
* Hence, **22.5% of patients are suffering from pneumonia** and the remaining **77.5% are pneumonia negative**.

****

**Action Taken:** Data is not balanced, to balance the data augmentation technique is used during data generation.

1. **Distribution of 3 different classes data with target:**

* The count of patients with *No Lung opacities/ Not normal* is higher than the pneumonic or normal patients
* the count of normal class is less than other 2 classes indicating that the data has a greater number of Ill health patients

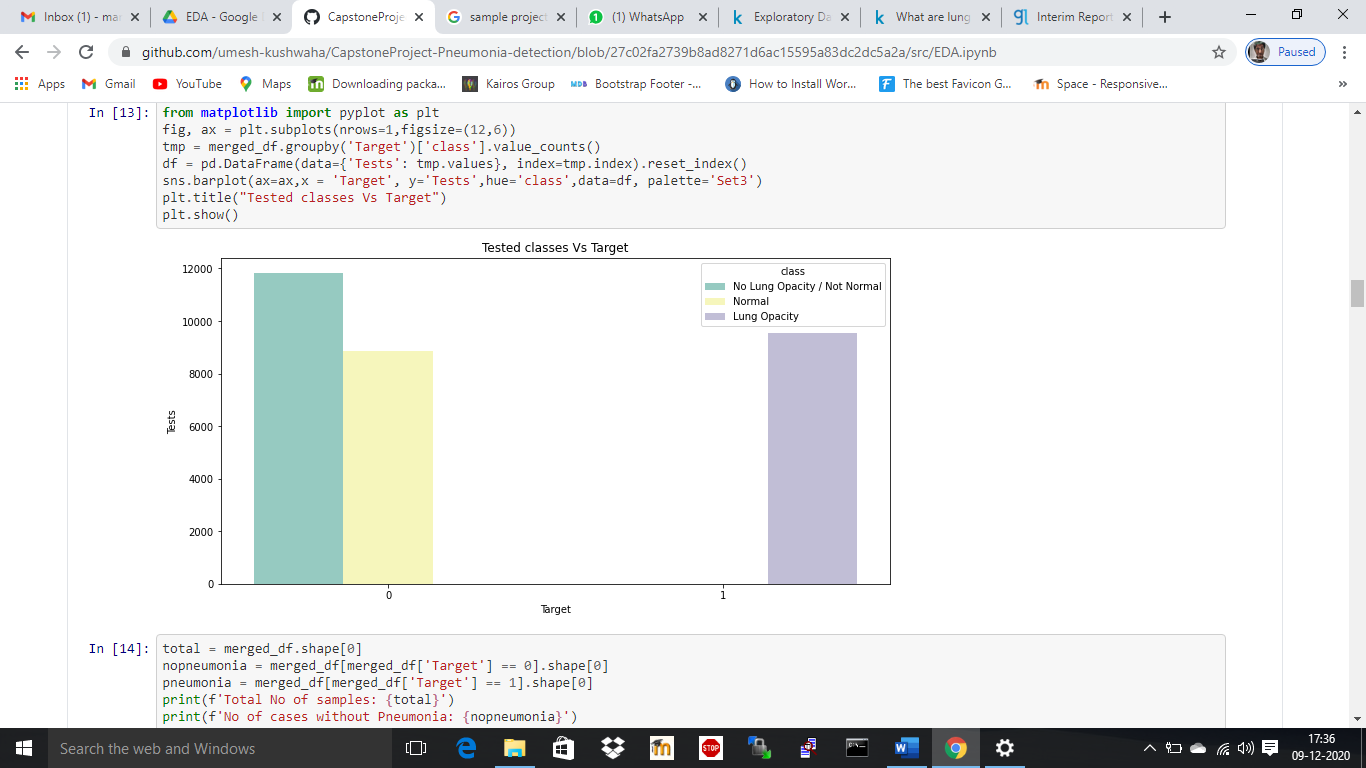


**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Distribution of target (Pneumonia & Non-Pneumonia) with 3 different classes:**

The same can be observed when plotting the count of Target values segregating the classes.

* Patients with *No Lung Opacity/ Not normal* observations are more than there other 2 classes



**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Train set:**

* 26,684 images are available in the training set are unique (equal to unique patient IDs).

1. **Bounding box:**

* Out of 26,684 images available in the training set, 2614 images have only 1 bounding box, 3,266 images have 2 bounding boxes, 119 images have 3 boxes, 13 images have 4 boxes and 20672 images do not contain lung opacities.
* 3,398 patients have more than 1 bounding box.
* If any patient has at least one lung opacity area, then the patient is considered to have pneumonia.

|  |  |
| --- | --- |
| **No of Occurrences** | **Count of the Patient ID** |
| 0 | 20672 |
| 1 | 2614 |
| 2 | 3266 |
| 3 | 119 |
| 4 | 13 |

1. **Data characteristics:**

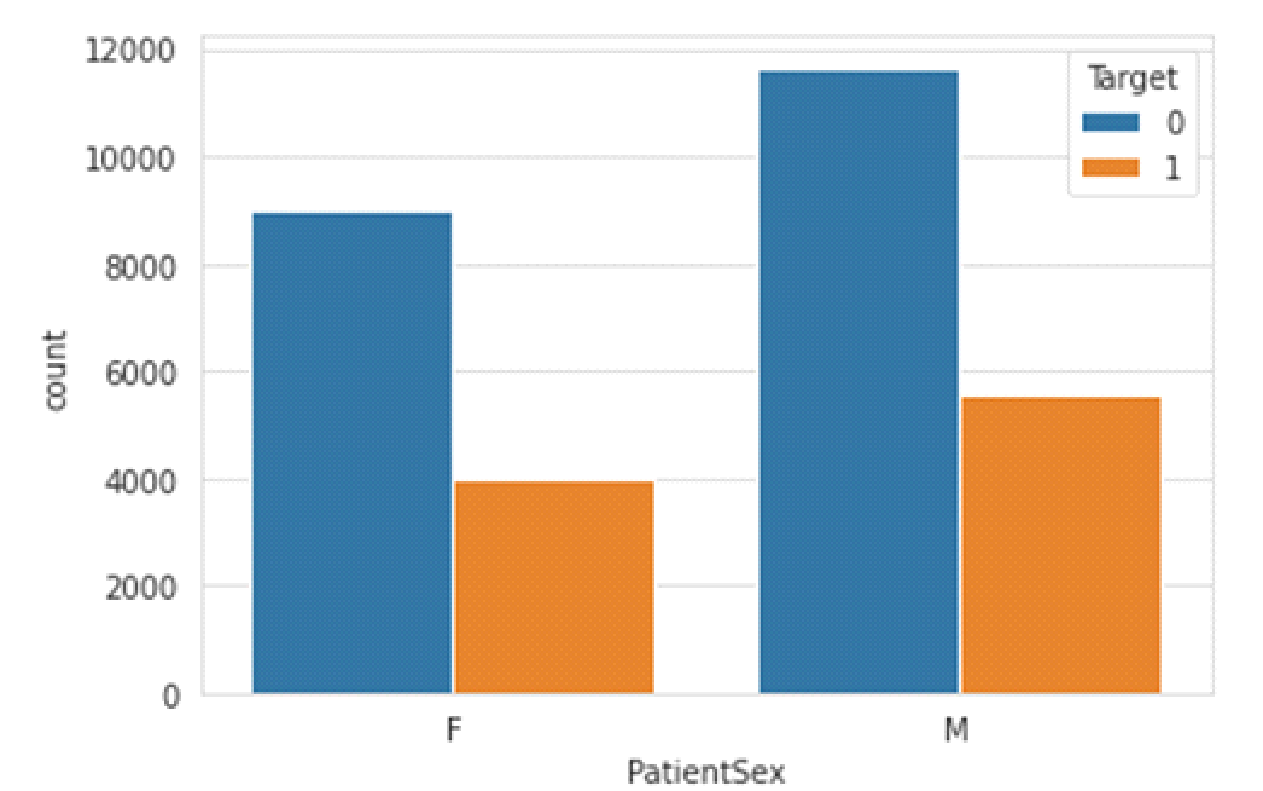
* We have different parameters or characteristics of available information – patient age, sex, body part examined, view position, rows and columns, pixel spacing, etc.

1. **Correlation:**

* We have observed that ‘Target’ and ‘View Position’ have a higher correlation and stand at 0.42.

1. **Gender mix: Distribution Gender Vs Target information**

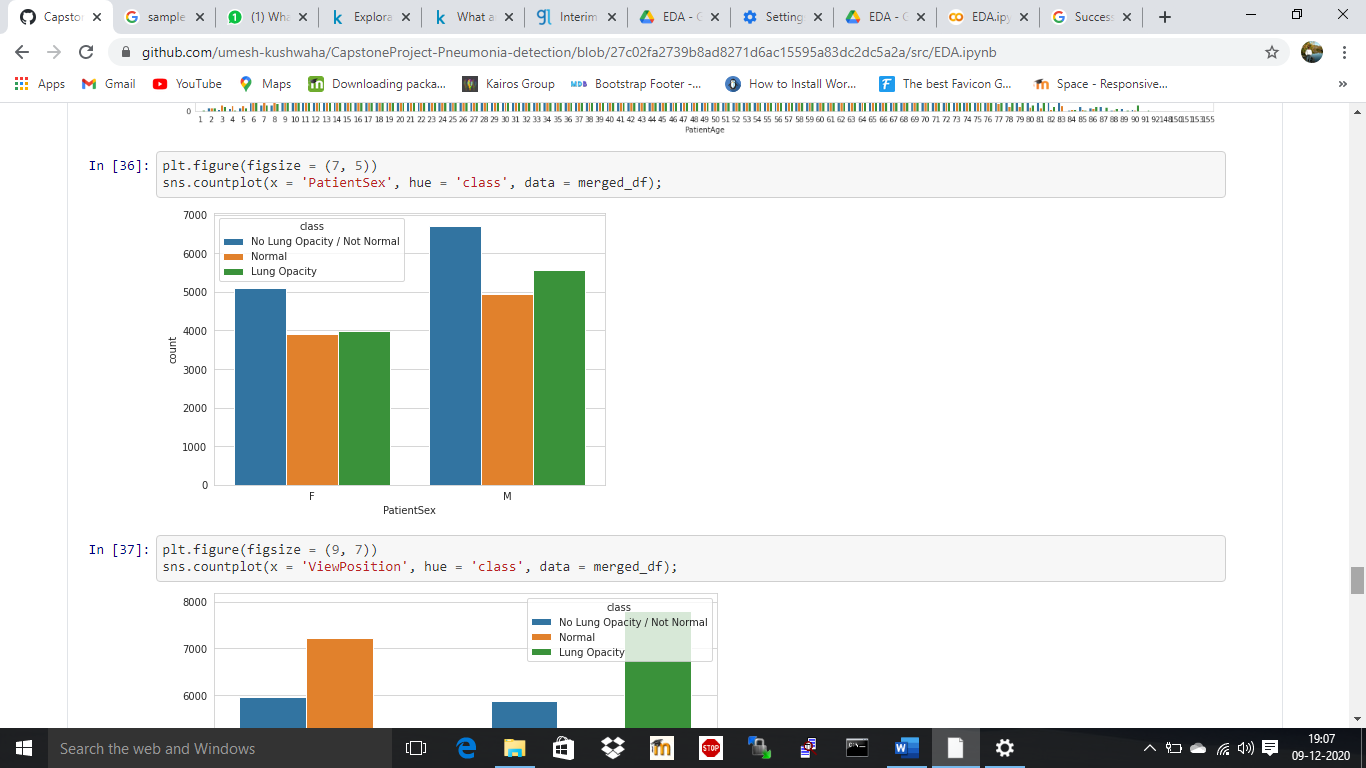
* Looking at the high pillars, there are more test samples (radiographs) for males than females.
* This could be due to, Men consume more alcohol, work outdoor more than women, smoke.
* Out of total 9,555 cases of Lung opacity, ~60% is male and rest 40% is female.
* Approximately, one third of the total cases are diagnosed as pneumonia for both the genders.



**Action Taken**: To balance the data augmentation with random shuffling is used.

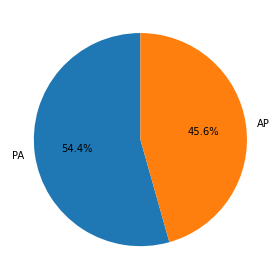
1. **Gender mix: Distribution Gender Vs class information**

* No opacity but Not Normal cases constitute higher number indicating that the patients could be suffering from other lung related illness but pneumonia.
* Males who are diagnosed for lung opacities (Pneumonia) are slightly high in number comparatively.

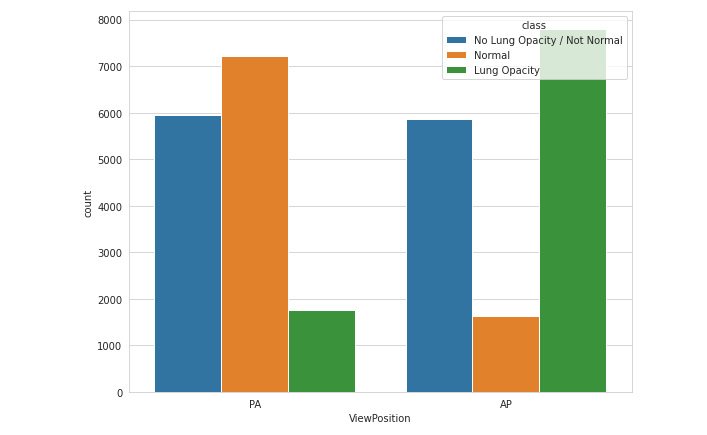


**Action Taken**: To balance the data augmentation with random shuffling is used.

1. **Distribution of View position:**



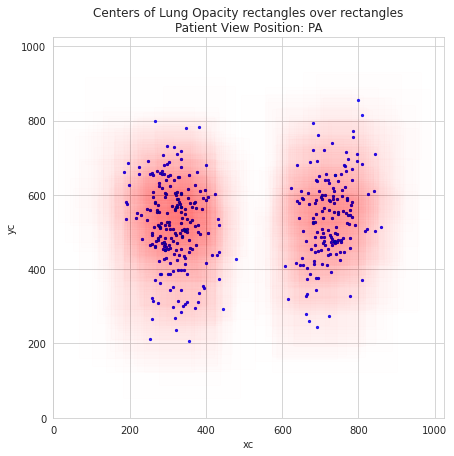
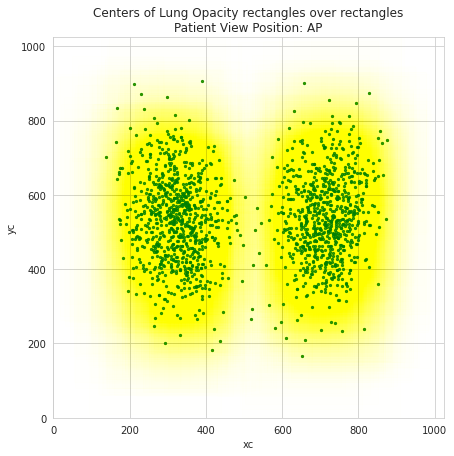
* Radiographs images with PA position are considered to be of good standard in medical profession. The PA amounts to 54.4% of total radiograph images. Therefore, data augmentation with random shuffling to balance the data to spread the impact of AP evenly.



* Evidently **PA position pointing considerably less lung opacities** than the AP position. Whereas the *no opacity but not normal* class seems to be the same in both the position.
* Scanning in AP position is usually done when the patient is not able to standup or for some reason the frontal scanning is not possible.
* This means a patient who has undergone scanning in AP position is more likely to be diagnosed for Pneumonia.
* If the same patient undergoes scanning in PA position, chances are that the results might come out normal.

**Action Taken**: Data augmentation with random sampling is used to balance the data. However, this is an important insight to be noted and taken care in the model improvement task.

* In the below chart, concentration of Lung opacities for AP is larger, whereas for the PA cases, it’s less.



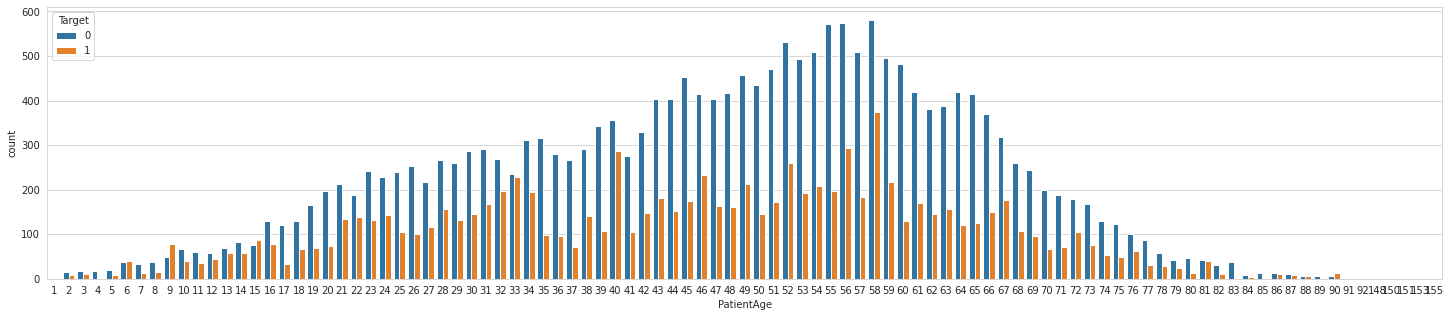
1. **Distribution data over different age group:**

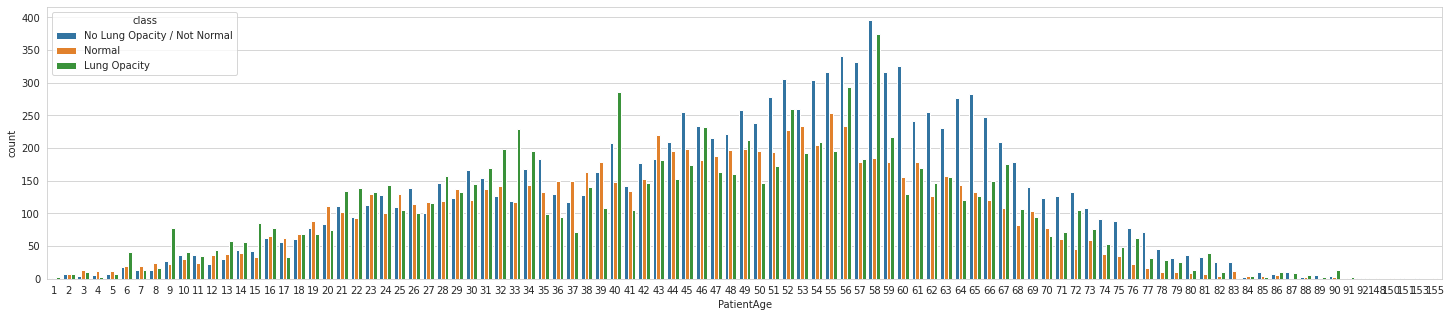
* Three forth of the total test reports falls within age group of 25 – 70. The peak is between 50-60. It can be observed that lung related ailments are more common in this age group.

This is possibly due to,

* In Adults, the tests conducted are more. Possibly due to adults work out door, consume alcohol, smoke, exercise less (or lazy or inactive than children), suffer from obesity, diabetes, etc. therefore the tests done are more. Not all test result attribute to pneumonia therefore the ratio between pneumonia and normal is more: therefore, the pillars in the below graph.
* In children, the tests for lung ailments conducted are considerably less because of unawareness of possibility of lung related ailments in young age. Therefore, the ratio between Pneumonia to Normal is very less.

* Pneumonia cases are higher between the age group of 30-65 years.
* Possibly due to adults work out door, consume alcohol, smoke, exercise less (or lazy or inactive than children), suffer from obesity, diabetes, etc.
* The lung opacities are spread across the patient age and have peaks within age 50 to 65 years.

****



**Action Taken**: Data augmentation with random sampling is used to balance the data

**Data Preprocessing:**

* **Image scale:** The images have been windowed and leveled already, as have been rescaled to 8-bit encoding and the resolution has been rescaled to (1024, 1024).
* **Data discrepancy:** There is no discrepancy in the data, as the data in the class csv and label csv is the same.
* **Data imbalance:** Data augmentation and random shuffling is used to balance the data.

Data Augmentation:

* For data augmentations, techniques such as image rotation, scaling, translating, changing brightness, contrast, blurring and sharpening are used.

**Deciding Models**

The goal is to predict whether the patient is suffering from Pneumonia or not, therefore there are 2 related tasks,

* The radiograph images have to analyzed for potential lung opacities
* Predicted if the lung opacities attribute to the pneumonia or not.

Keeping the goal in mind, the below steps are performed to choose a model

* Identify the model that are pretrained. This can reduce the training time for the data
* Check the accuracy and other attributes of the model
* If required Add or drop layers from the model to improve the performance of the model.

The below algorithms are taken into consideration for building the model

1. VGG - Very Deep Convolutional Networks for Large-Scale Image Recognition. VGG has achieved 92.7% top-5 test accuracy in ImageNet, for a dataset of over 14 million images belonging to 1000 classes
2. Inception – A network known for its speed and accuracy. Due to the large number of training images (~26K, the network is used as backbone network to find the bounding boxes.
3. Mask RCNN - A deep neural network aimed to solve instance segmentation problem

All the three models are trained and tested with the available datasets by preprocessing the data. With the results Mask RCNN was chosen as the appropriate algorithm as it predicts the Lung Opacities at the same time it also arrives at the prediction that if that patient has pneumonia or not.

**Model Building**

Mask RCNNcan separate different objects in an image or a video. You give it an image, it gives you the object bounding boxes, classes and masks. The model is built as below

* All the required libraries are imported and the required data sets are included
* Introduce image augmentation techniques to increase the number of training images as identified in EDA. New images are introduced in the training dataset by
  + Changing the data geometrically by scaling and transforming the images
  + Changing the brightness and contrast of the images
  + Blurring or sharpening the Images
* Configure the model by introducing RESNET50 as the backbone model
* Include the GPUs used for the training the model and also the number of images used per GPUs
* Configure the number of classes required to be detected by the model
* Start training the model by including the MaskRCNN model with the configured data
* Load the weights for the model and start training the model for 30 Epochs.

**Model Performance**:

Accuracy: 78.36%

loss: 1.2235

mrcnn\_class\_loss: 0.1848

mrcnn\_bbox\_loss: 0.3528

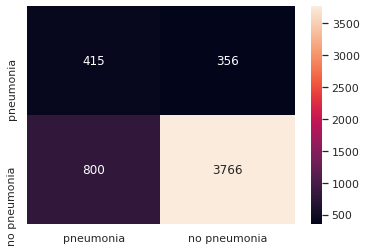
mrcnn\_mask\_loss: 0.3628

val\_loss: 1.2829

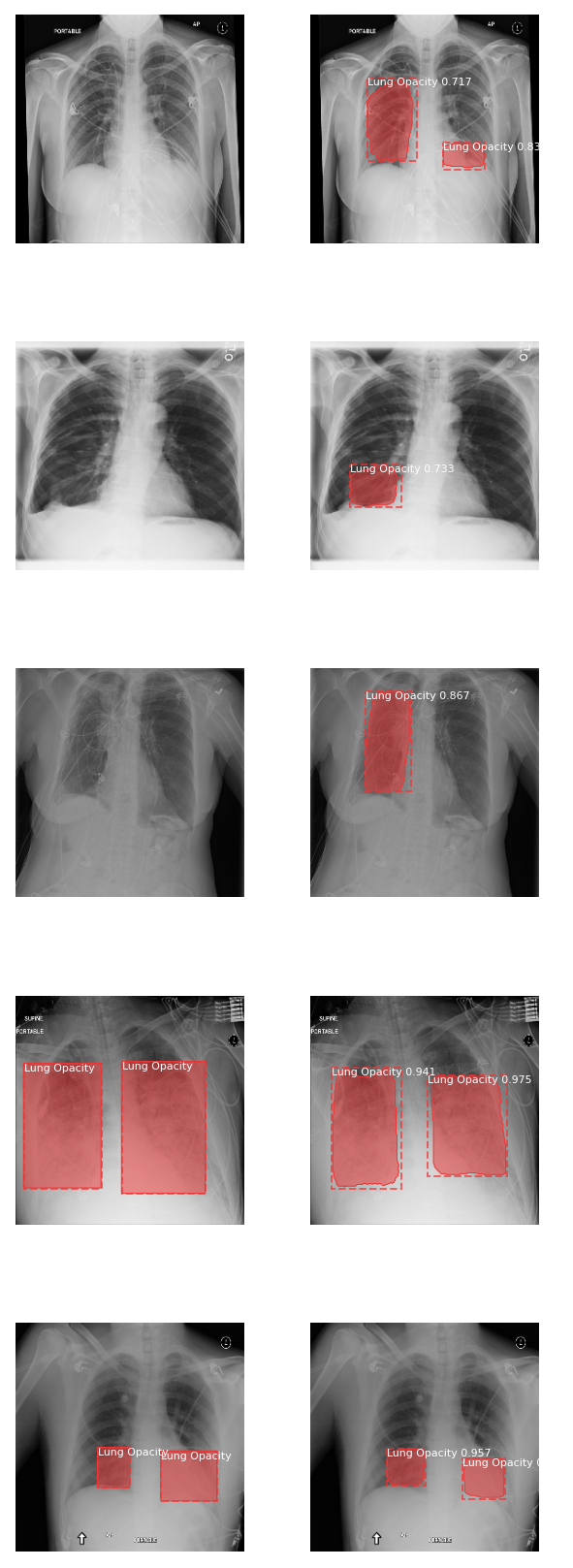
val\_mrcnn\_class\_loss: 0.19

val\_mrcnn\_bbox\_loss: 0.38

val\_mrcnn\_mask\_loss: 0.37



Attached a screenshot which indicates bounding boxes in the lung radiograph images.



**Appendix**

**Code:**

|  |  |  |
| --- | --- | --- |
| **1** | Mask RCNN Model |  |
| **2** | VGG16 based Model |  |
| **3** | Inception based Model |  |